Predicting Heart Disease Risk Using Machine Learning

Vladyslav Shashkov, Yann Ravot, Andronikos Loukaidis

1 Introduction

The model has 91% predictions of -1's and 9% predictions of 1's, reflecting the dataset where approximately 1/10 patients is sick. We perform 100 iterations. If we take a too large step-size so that the model still converges, we risk to overfit, as the number of -1's is much greater than the number of 1's..

Methodology

Data Processing

Data preprocessing involved handling missing values, scaling features, and removing outliers. For each feature, missing values were replaced with the mean, and feature scaling was performed to normalize the data.

Models Implemented

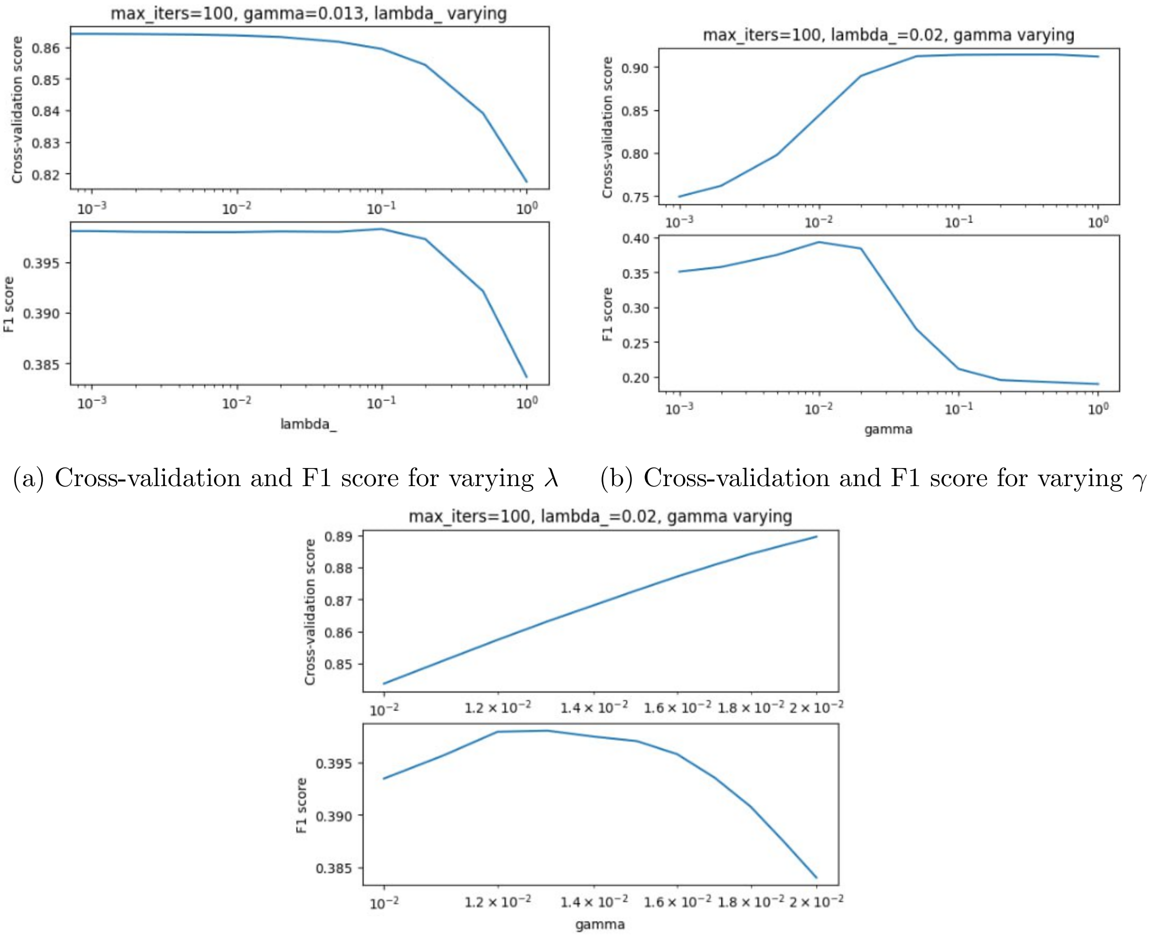
The following machine learning models were implemented:

- **Linear Regression (GD and SGD)** - Linear models using gradient descent and stochastic gradient descent.
- **Least Squares and Ridge Regression** - Regression methods using normal equations, with ridge regression incorporating regularization.
- **Logistic Regression and Regularized Logistic Regression** - Classification models optimized via gradient descent, with regularization in the latter to prevent over-fitting.

Hyperparameter Tuning

Hyperparameters, particularly the learning rate γ and the regularization term λ, were tuned using cross-validation. This process helped to optimize model performance while avoiding overfitting.

Results



(a) Cross-validation and F1 score for varying λ

(b) Cross-validation and F1 score for varying γ

(c) Cross-validation and F1 score for fine-tuned γ range

Figure 1: Hyperparameter Tuning Results

Figure 1a shows the effect of varying λ on cross-validation and F1 scores. As λ increases, the model's performance decreases, indicating that a low regularization is preferable.

In Figure 1b, the learning rate γ is varied. The cross-validation score peaks at moderate values of γ, after which it starts to decline, likely due to instability in gradient descent.

Figure 1c provides a closer look at γ values around the optimal range. A peak in F1 score is observed, reinforcing the choice of $\gamma = 0.013$ for best performance.

Discussion

The results indicate that regularized logistic regression with a low λ and a learning rate around 0.013 provides the best performance. This balance reduces overfitting while maintaining a high prediction accuracy.

Conclusion

This project successfully demonstrates the application of machine learning methods to predict heart disease risk. Logistic regression with regularization and a tuned learning rate achieved the best performance. Future improvements may include using external libraries for more sophisticated models.